EXTENDING OUTPUT ATTENTIONS IN RECURRENT NEURAL NETWORKS FOR DIALOG GENERATION

Chanseung Lee

Lane College
Eugene, OR USA.

ABSTRACT

In natural language processing, attention mechanism in neural networks are widely utilized. In this paper, the research team explore a new mechanism of extending output attention in recurrent neural networks for dialog systems. The new attention method was compared with the current method in generating dialog sentence using a real dataset. Our architecture exhibits several attractive properties such as better handle long sequences and, it could generate more reasonable replies in many cases.

KEYWORDS

Deep learning; Dialog Generation; Recurrent Neural Networks; Attention

1. INTRODUCTION

In language dialogue generation tasks, people have heavily used statistical approach which depends on hand-crafted rules. However, this approach prevents the system from training on the large human conversational corpora, which become more available by day. Consequently, it was only natural to invest our resources towards the development of data-driven methods and generating novel natural language dialogs.

Amongst data driven language generation models, recurrent neural networks (RNN), long short-term memory [1] and gated recurrent [2] neural networks in particular, have been widely used in dialog systems [3] and machine translation [4]. In RNN-based dialog generative models, the meaning of a sentence is mapped into a fixed-length vector representation and then they generate a translation based on that vector. There was no need to rely on n-gram counts or the capture of text's high-level meaning, since the systems generalize to new sentences superior to other approaches.

One of the recent advancement in sequence learning is attention mechanism. By the usage of attention mechanism, the full encoding of source sentence into a fixed-length vector is no longer necessary. Instead, researchers enable the decoder to "attend" the various segments of the source sentence at each step of the output generation process. It has been proven to be that neural attention can be used for various sentence-to-sentence tasks with excellent results. It associates salient items amongst the source sequence with the items generated in the selected target sequence [5] [6].

In this paper, the research team introduce a new attention method, called 'output attention', in a recurrent neural network (RNN) language model. Throughout each and every step of a RNN, the function for computing the next output state collects the weighted average of all the previous

DOI : 10.5121/ijaia.2018.9504

outputs. The basic idea of output attention is that when you generate next word, the previous words you already generated are clearly important in choosing next word. Therefore, researchers assign different weights to each of the previous output and their combined attention value is fed into the current output function. With output attention mechanism, the network can take the outputs produced many time steps earlier into consideration.

2. RELATED WORK

Many work has been done in the area of statistical machine translation-based response generation [1] [7] [8] [9]. The generation of dialogue response as a statistical phase-based machine problem, which doesn't require explicit man-made rules, has been formulated by Ritter, Cherry, and Dolan [10]. RNN's recent success in the field of statistical machine translation [1] [7] has brought forth an inspiration for application of such models in the field of dialogue modeling.

Other researchers have recently used SEQ2SEQ method to directly generate responses in an end-to-end fashion without using SMT phrase tables [11] [12]. Vinyals and Le [11] and Shang, Lu, and Li [12] employ an RNN to generate responses in human-to-human conversations by treating the conversation history as one single sequentially ordered data. However, in such models, the distant relevant context in the dialog history is difficult to remember and some efforts have been made to overcome this limitation.

Sordoni et al. [9] encoded the most recent message and the previous context using a bag-of-words representation, which is then decoded using an RNN. This approach computes the distance of each word in the generated output to all the words in the conversation history. But this approach loses the temporal and relevant information of the conversation history.

Serban et al. [2] employ a hierarchical model that stacks an utterance-level RNN on a token level RNN. By doing so, the utterance-level RNN reduces the number of computational steps between utterances. Wen et al. [13] and Wen et al. [14] improve spoken dialog systems via multi-domain and semantically conditioned neural networks on dialog act representations and explicit slot value formulations.

Recently, Tran, Bisazza, and Monz [15] demonstrated that the memory network mechanism can improve the effectiveness of the neural language model.
In this paper, we propose a new attention-based neural language model for dialogue modeling that learns more of the past relevant information in a conversation history, and thus generate more reasonable responses in the dialog.

3. MODEL ARCHITECTURE

In this section, researchers describe the architecture of the new extended output attention model and discuss how these additional attentions can help to achieve efficient dialog generation for sequence-to-sequence learning.

Encoder: A recurrent neural network is a neural network that consists of a hidden state $h$and an optional output y which operates on a variable-length sequence $X = \{x_1, x_N\}$. At each time step t, the hidden state h_t of RNN is updated by

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

Where f is a non-linear activation function. f is the nonlinear function in the recurrent unit, which can be implemented in a non-linear activation function, or Long Short-Term Memory (LSTM) [3], or Gated Recurrent Unit (GRU) [4].

Decoder: In traditional model architecture, define each conditional probability as follows

$$p(y_t | y_1, \cdots, y_{t-1}, X) = g(y_{t-1}, s_t, c_t) \tag{2}$$

where s_t is an RNN hidden state for time t, computed by

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{3}$$

The probability is conditioned on a traditional attention vector c_t for each target word y_t. This implements a mechanism of attention in the traditional decoder.

By using attention, the decoder is now able to decide which parts of the source sentence to pay Attention to. In addition, by letting the decoder have an attention mechanism, the encoder is now relieved from the burden of having to encode all information in the source sentence into a fixed length vector.

The context vector c_t depends on a sequence of annotations $(h_1,...,h_t)$ to which an encoder maps the input sentence. Each annotation h_i contains information about the whole input sequence with a strong focus on the parts surrounding the i-th word of the input sequence. The context vector c_t is, then, computed as a weighted sum of these annotations h_i:

$$c_t = \sum_{j=1}^{t} \alpha_{tj} h_j \tag{4}$$

The weight α_{tj} of each annotation h_j is computed by

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})} \tag{5}$$

Where,

$$e_{tj} = a(h_j, h_t) \tag{6}$$

represents a score of how well the inputs around position j and the output at position t match.

The probability α_i reflects the importance of the annotation h_i with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i. Figure1 shows the structure of traditional attention method.

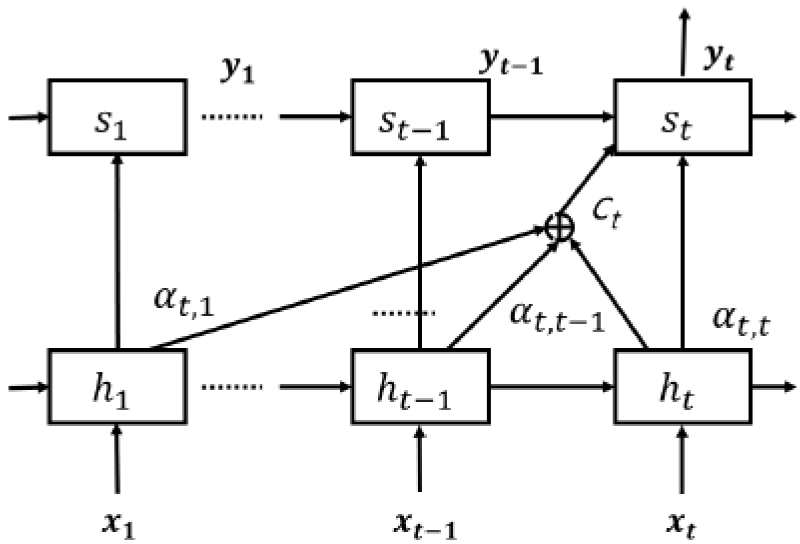

Figure 1 Architecture of traditional attention

In this paper, the team of researchers modify the recurrence formula by adding a new output attention vector d_t to the input of the LSTM. As for the following conditional probability, Eq. (2), the term s_t in Eq. (3) is now modified as follows

$$s_t = f(s_{t-1}, d_{t-1}, c_t) \tag{7}$$

The main difference between Eq. (3) and Eq. (7) is that Eq. (7) uses d_{t-1} instead of y_{t-1}.

While y_{t-1} is a single output value of previous time step, d_{t-1} is an *output attention* vector which represents the weights of each output already generated. By taking previous outputs into consideration, the formula of d_t is defined as

$$d_t = \sum_{j=1}^{t} \beta_{tj}\, y_j \tag{8}$$

$$\beta_{tj} = \frac{\exp(f_{tj})}{\sum_k \exp(f_{tk})} \tag{9}$$

The probability β_i reflects the importance of the annotation h_i with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i.

$$f_{tj} = b(s_j, h_t) \tag{10}$$

The definition of c_t, α_{tj}, e_{tj} remain unchanged. Figure2 shows the structure of the proposed model in this paper.

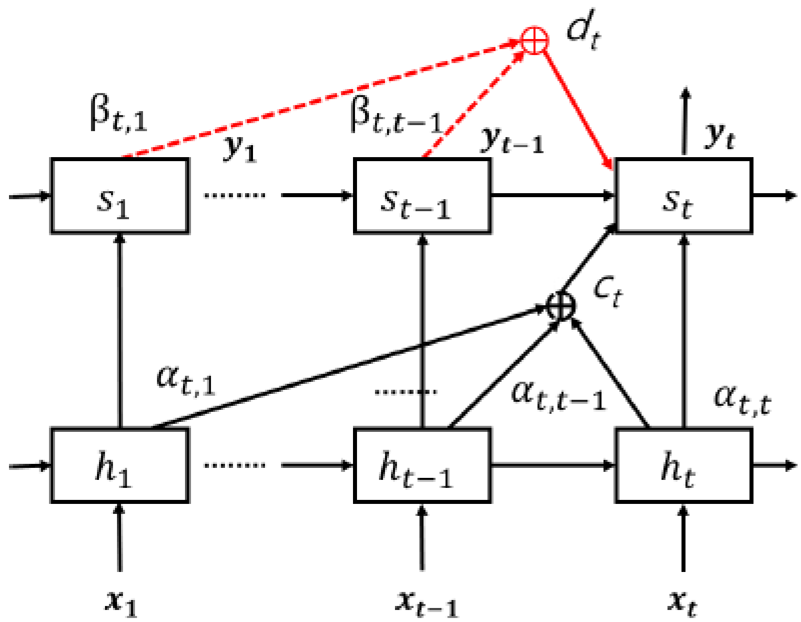

Figure 2 Architecture of extended output attention

In Figure 2, d_t represents *output attention* vector, and is a new extended attention vector for output values.

4. Experimental Results

In this section, researchers describe the experimental results with some datasets and show some samples of the interactions with the system that we trained. They used Movie Dialog Data for the experiment. This dataset contains a metadata-rich collection of fictional conversations extracted from raw movie scripts. It contains 220,579 conversational exchanges between 10,292 pairs of movie characters and total of from 617 movies total 304,713 utterances from 9,035 characters.

Researchers trained a single layer LSTM with1024 memory cells using stochastic gradient descent with gradient clipping. The vocabulary consists of the most common 20Kwords, which includes special tokens indicating turn taking and actor. Below is a few samples of simulated dialog sessions. Table 1 shows some sample utterance between machine and human. In Table 1, 'Input'is given by human and 'Output attention' is the response generated by the proposed system in this paper and 'Attention' is the response from traditional attention method.

Researchers find it encouraging that the model can better understand contexts, and generate more reasonable sentences compared to the traditional model. The model does so without any explicit knowledge representation component except for the additional output attention.

One drawback of this basic model is that it only gives simple, short, and sometimes unsatisfying answers to our questions as can be seen in Table 1. Indeed, if a researcher asks not identical but semantically similar questions, the answers can sometimes be inconsistent. This is expected due to the simplicity of our model and the dataset in our experiments.

Table 1 Summary of dialog

Input	**Output attention**	**Attention**
How do you do?	i am always bad	i am not this.
are you religious ?	we're all like that	we're all here for you
i'm thirty four .	i'm twenty years old. how old are you ?	i'm twenty years old.
all right i never should have bought them !	they're killing me !	they're killing me !
what's so interesting about him ?	i'm just saying that	i'm just on vacation
i can't blame anybody for anything i do	we're partners	we're gonna ?
much more terry. much much more !	i'm trying my best	i'm trying to find out with
they said she was here	she's not at home ?	she's not a ? she's just
sure you can if you want to.	i'm not sure i can	i'm not sure i can i . to
bank teller is just a start. i got plans.	you're talking about giving up baseball ?	you're talking about a ?
i am not smiling	i am always down	i am not a loser
who's leaving ?	we're leaving	we leaving
are you alright ?	i am good. thanks how are you	i am good
i m not even sure why	i m not sure .	i m so honey

5. CONCLUSIONS

In this paper, researchers propose a new output attention mechanism to improve the coherence of neural dialogue language models. The new attention allows each generated word to choose which related word it wants to align to in the increasing conversation history. The results of our experiments show that it's possible to generate simple yet coherent conversations using extended attention mechanism. Even though the dialog the model generates has obvious limitations, it is clear that another layer of output attention allows the model to generate more meaningful responses.

For future work, the model may need substantial enhancements in many aspects in order to convey even more realistic dialogs. One possible direction of future work is to employ memory network method to facilitate remembering long term history in the dialog.

REFERENCES

[1] Hochreiter S1 and Schmidhuber J., Long short-term memory, Neural Computation, 1997 Nov. 15;9(8):1735-80.

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, EMNLP 2014.

[3] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, Joelle Pineau, Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models, AAAI 2016.

[4] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR 2015

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,ICML 2015

[6] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, Jakob Uszkoreit, A Decomposable Attention Model for Natural Language Inference, EMNLP 2016

[7] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to Sequence Learning with Neural Networks, NIPS 2014

[8] Minh-Thang Luong, Hieu Pham, Christopher D. Manning, Effective Approaches to Attention-based Neural Machine Translation, EMNLP 2015

[9] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob G. Simonsen, Jian-Yun Nie,A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion, Conference of Information Knowledge and Management (CIKM) 2015

[10] Alan Ritter, Colin Cherry, and William B. Dolan, Data-Driven Response Generation in Social Media, EMNLP 2011

[11] Oriol Vinyals, Quoc Le, A Neural Conversational Model, ICML Deep Learning Workshop 2015

[12] Lifeng Shang, Zhengdong Lu, Hang Li, Neural Responding Machine for Short-Text Conversation, ACL 2015

[13] Tsung-Hsien Wen, Milica Gaˇsi´c, Nikola Mrkˇsi´c, Pei-Hao Su, David Vandyke and Steve Young , Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems, EMNLP 2015

[14] Tsung-Hsien Wen, Milica Gaˇsi´c, Nikola Mrkˇsi´c, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes,David Vandyke, Steve Young, Conditional Generation and Snapshot Learning in Neural Dialogue Systems, arXiv:1606.03352, 2016

[15] Ke Tran, Arianna Bisazza, Christof Monz, Recurrent Memory Networks for Language Modeling, NAACL 2016